A Data Analysis Approach

Where to put Headquarter?

Toronto vs. New York

Bo-Xiang Wang

IBM Data Science Professional Certification

Capstone Project

April, 2020

Introduction

This project will analyze neighborhoods between Toronto, Canada and New York City, New York. A Fortune 500 company is looking to move its headquarters to either Toronto or New York City. The company wants insight into the neighborhoods and local businesses in the cities so that its employees may have the optimum living standards and quality of life. This project will explore the similarities and dissimilarities between certain neighborhoods in the two cities, and determine which neighborhoods best fit the culture of the Fortune 500 company’ s employees.

Data

The data used for this project will be acquired from the respective cities Wikipedia website pages. The datasets consist of the postal codes, neighborhood names, latitude, and longitude information for each neighborhood. Foursquare API search feature will be used to collect neighborhood venue information. Details about local venues and locality will be provide insight into the qualities of a neighborhood. In addition to Foursquare, various python packages will be used to create maps and machine learning models to further provide insights into our neighborhood battle project.

The following datasets from these websites:

- Toronto Neighborhoods- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
- Toronto Latitude and Longitude - http://cocl.us/Geospatial_data
- New York City neighborhoods - https://geo.nyu.edu/catalog/nyu_2451_34572
- New York City Latitude and Longitude = Python Geolibrar

Methodology

Work Flow

1. HTTP requests would be made to this Foursquare API server using zip codes of the Seattle city neighborhoods to pull the location information (Latitude and Longitude).
2. Foursquare API search feature would be enabled to collect the nearby places of the neighborhoods. Due to http request limitations, the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 700.
3. Folium- Python visualization library would be used to visualize the neighborhoods cluster distribution of Seattle city over an interactive leaflet map.
4. Extensive comparative analysis of two randomly picked neighborhoods world be carried out to derive the desirable insights from the outcomes using python’s scientific libraries Pandas, NumPy and Scikit-learn.
5. Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two chosen neighborhoods would be analyzed individually collectively and comparatively to derive the conclusions.

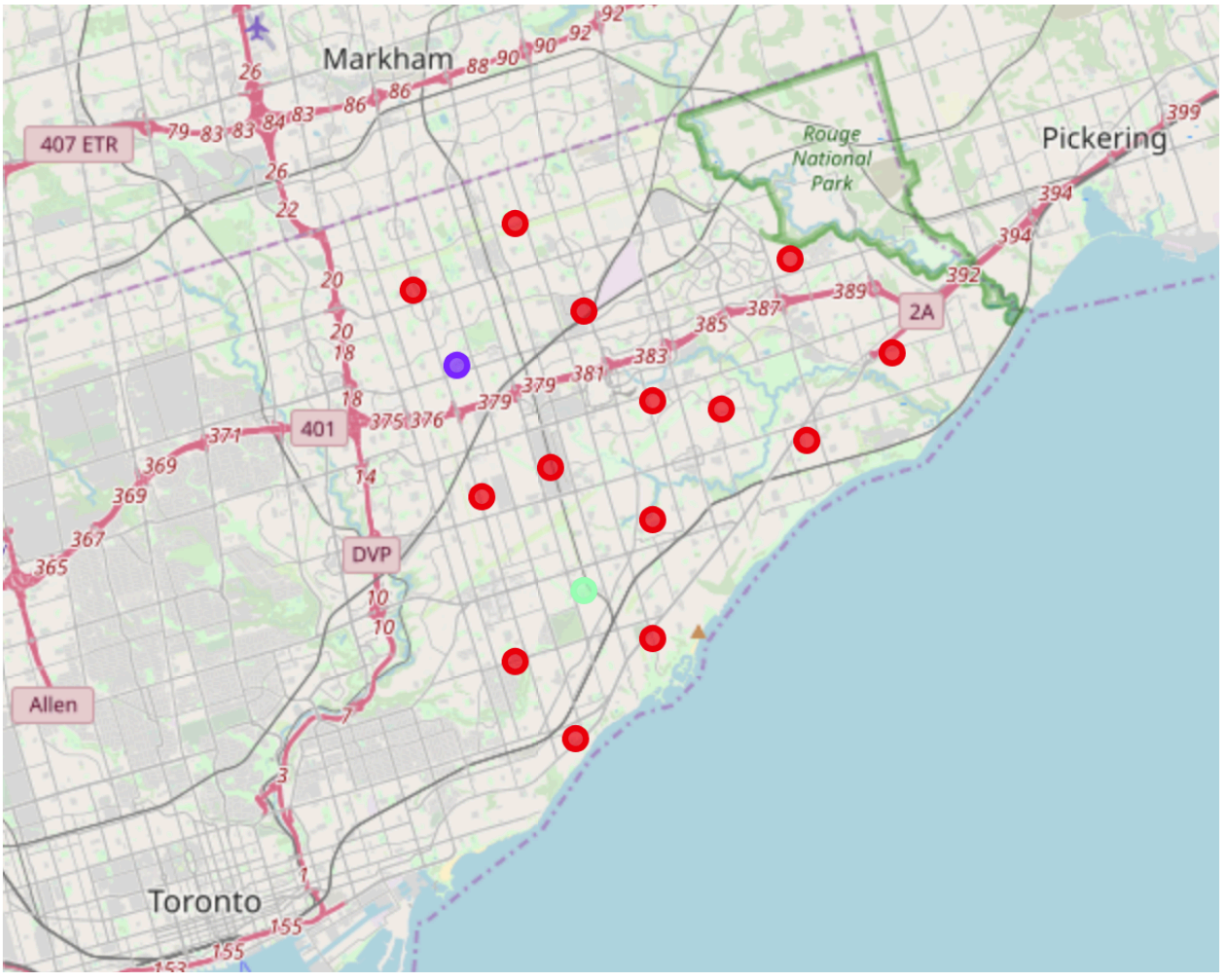
The following are the Python packages

- Pandas - Library for Data Analysis
- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

Results

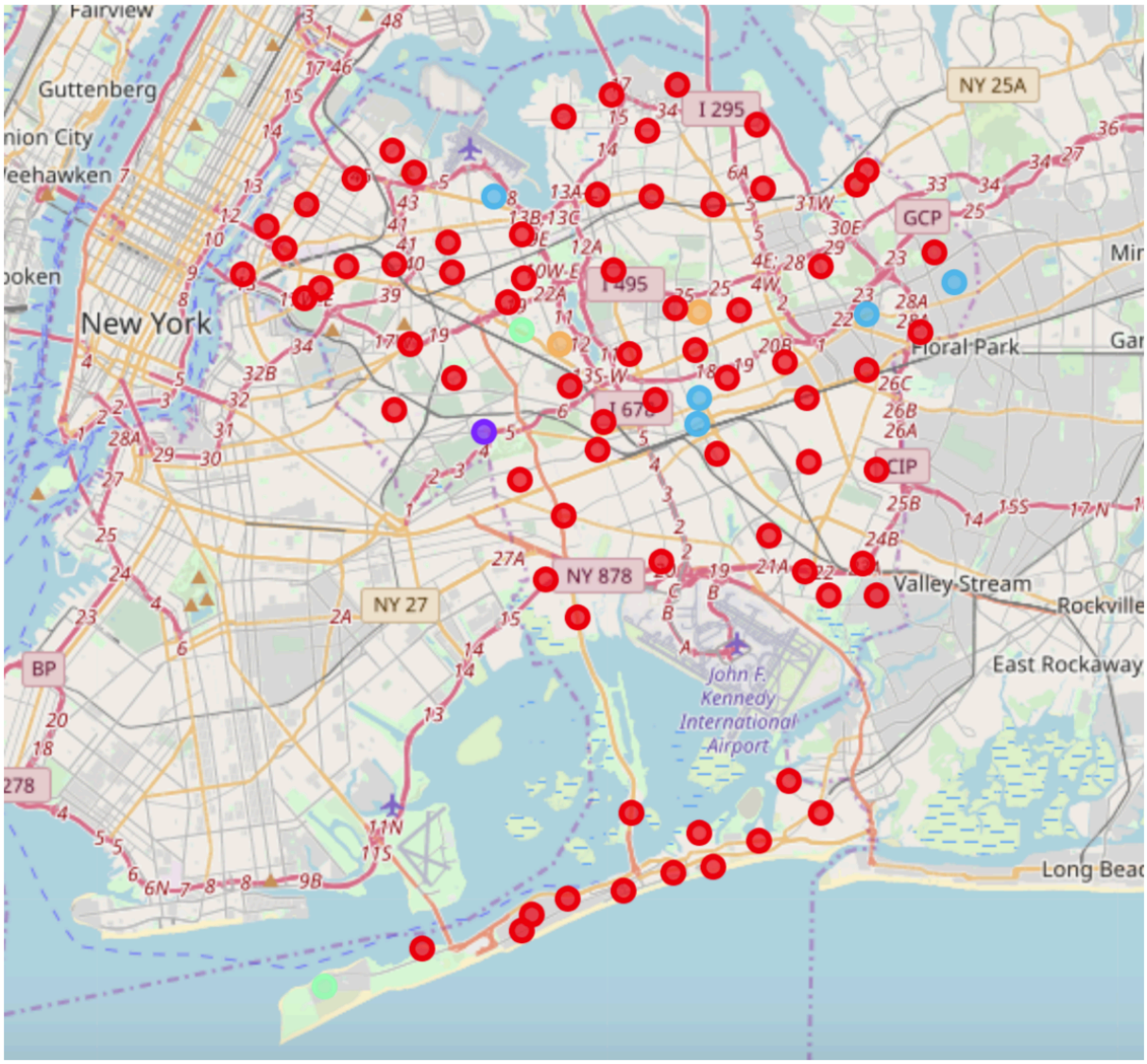
Scarborough Borough in Toronto, Canada

I use k-means to group the neighborhoods in Scarborough into 3 clusters. Cluster_0 has 15 neighborhoods and the most common venues are skating rinks, international cuisine restaurants and breakfast spots. Cluster 1 has 1 neighborhood 1 neighborhood, and the most common venues are pizza place and noodle house. Cluster 2 has 1 neighborhood, and the most common venues are Chinese restaurants and discount stores.



Queens Borough in New York City

I used k-means to group the Queens borough into 5 clusters. Cluster_0 has 81 neighborhoods and consist of many international cuisine restaurants and grocery stores. The most common venues are pizza places, deli, and Chinese restaurants. Cluster_1 has 1 neighborhood and the most common venue is a dance studio. Cluster_2 has 5 neighborhoods and the most common venue are donut shops and international cuisine restaurants. Cluster_3 has 2 neighborhoods and the most common venues are the beach and a bakery. Cluster_4 has 2 neighborhoods and the most common venues are gyms and donut shops.



Discussion

Toronto has 11 boroughs and 103 neighborhoods. The geographical coordinate of Toronto, Canada are 43.7170226, -79.4197830350134. In Scarborough borough, found 85 venues in 17 neighborhoods In Scarborough borough, the neighborhoods with the most venues are L’Amoreaux West and Steeles West.

There are 79 distinct venues in 50 categories. New York City has 5 boroughs and 306 neighborhoods. The geographical coordinate of New York City are 40.7308619, -73.9871558. Foursquare found 2108 venues in 81 neighborhoods in Queens borough. Many of the neighborhoods are homogenous and are very similar to each other. Both Scarborough and Queens borough consist of neighborhood cluster that contain majority of the neighborhoods, and the remaining cluster had 1-5 neighborhoods. Queens borough had a significant more number of neighborhoods and venues than Scarborough.

Conclusion

In conclusion, based on the quantity of venues and variety of venues, I would choose Queens over Scarborough as a choice to relocate the headquarters of the Fortune 500 company. Queens offer way more in choices for restaurants, gyms, grocery stores, and extracurricular activities for individuals and families of the company’s employees.